Multimedia Department

Od rozpoznawania do tłumaczenia mowy polskiej

Krzysztof Marasek
PJWSTK

Plan prezentacji

- ✓ Wprowadzenie
- ✓ Komponenty systemu tłumaczącego
- ✓ Aktualne działania

Dlaczego automatyczne tłumaczenie mowy?

- ✓ Czynniki techniczne
  - ✓ Choć nadal dalekie od perfekcji, metody maszynowego tłumaczenia dokonały w ostatnich latach wyraźnego postępu

    NIST MT workshops (od 2002)

    IWSLT workshops (od 2002)

    TC-STAR workshop (2005-2006)

    ACL/NAACL Shared Tasks (2005-2006)

TASK NAME	CATEGORY	SCOPE AND SCENARIOS
TIMIT	ASR	English phonetic recognition; small vocabulary
WSJ 0&1	ASR	Mid-to-large vocabulary; dictation speech
AURORA	ASR	Small vocabulary, under noisy acoustic environments
DARPA EARS	ASR	Large vocabulary, broadcast, and conversational speech
NIST MT OPEN EVAL	MT	Large scale MT, newswire, and Web text
WMT (EUROPARL/EUROMATRIX)	MT	Large scale MT, newswire, and political text
C-STAR	ST	Limited domain spontaneous speech
DARPA TRANSTAC	ST	Limited domain spontaneous dialog
IWSLT	ST	Limited domain dialog and unlimited free-style talk
TC-STAR	ST	Broadcast speech, political speech
DARPA GALE	ST	Broadcast speech, conversational speech

He, Deng, 11

    - ✓ Czynniki sukcesu to:

      Coraz silniejsze komputery (trening na klastrach PC, dużo RAM)

      Rozwój metod statystycznych i bazujących na danych (*data driven*)

      Odpowiednie zasoby lingwistyczne (korpusy bilingualne)

      Rozwój metod oceny jakości tłumaczenia (miary błędów)

      Stopniowe komplikowanie zadań

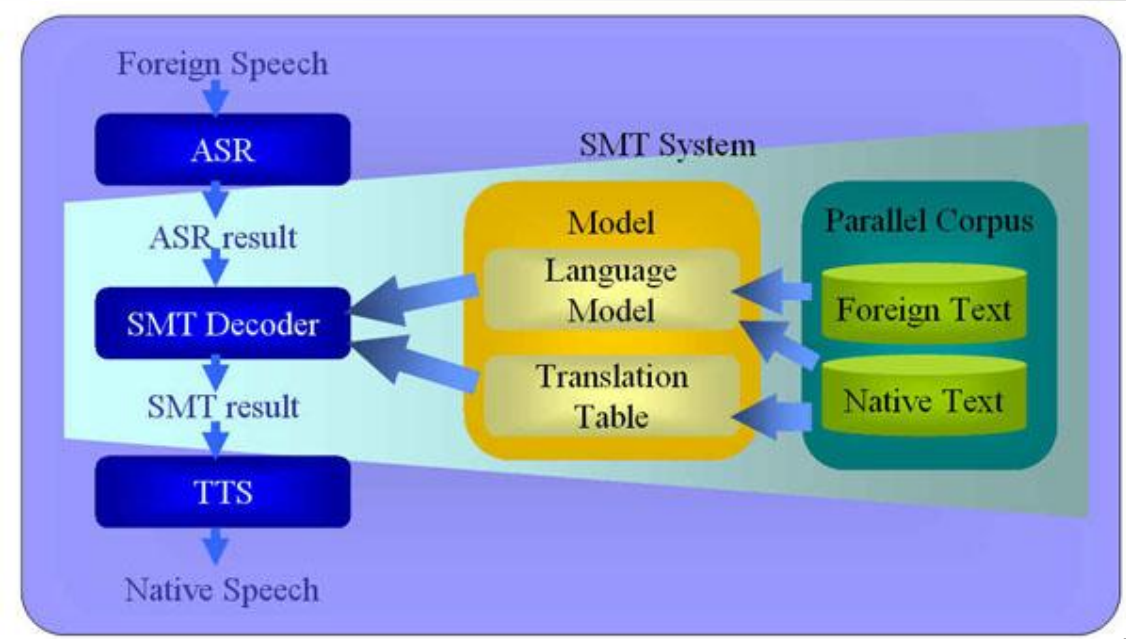
      Zaangażowanie wielkich firm: Google, Microsoft, IBM, itd..

Lazzari, 06

  - ✓ Rozwój interfejsów głosowych
    - ✓ Rozpoznawanie mowy działa coraz lepiej
    - ✓ Mowę syntetyczną coraz trudniej odróżnić od naturalnej
- ✓ Czynniki ludzkie
  - ✓ Potrzeby komunikacyjne w kurczącym się geograficznie świecie, komercyjne zastosowania SST

Komponenty systemu tłumaczącego

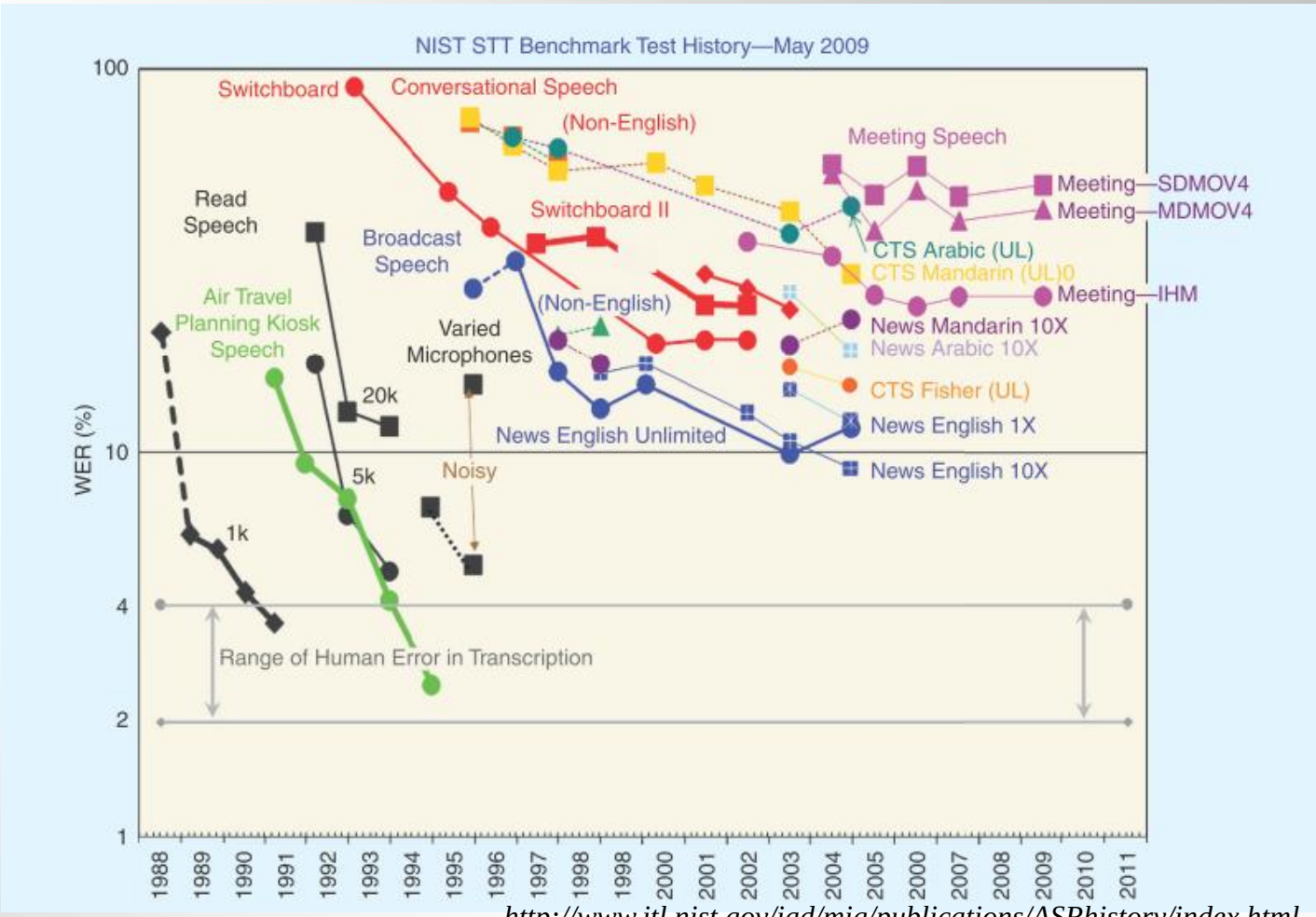
- ✓ Rozpoznawanie mowy
- ✓ Tłumaczenie wykorzystujące modele statystyczne



Intelligent Software Lab, 06

- ✓ Synteza mowy

Jakość rozpoznawania mowy



http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html

- ✓ **ASR da się używać!**
- ✓ ASR działa najlepiej dla ściśle określonego mówcy, im mniejszy słownik tym lepiej, angielski lepiej niż inne języki
- ✓ Stopa błędów rozpoznawania mowy spontanicznej jest co najmniej dwukrotnie większa niż dla czytania, stopa błędów jest wysoka dla konwersacji wielu mówców w trudnym akustycznie środowisku

Nasze aktualne wyniki ASR

- ✓ Projekt badawczy NCN N516 519439 System automatycznej transliteracji nagrań mowy polskiej – nagrania posiedzeń i komisji Senatu RP

$$\%\text{Correct} = \frac{H}{N} \times 100\%$$

$$\text{Accuracy} = \frac{H - I}{N} \times 100\%$$

Słownik w weryfikacji

Dekoder	ACC	CORR	PP
Julius + IRSTLM 44k	48	57	

$$Perplexity = 2^{-\sum_{i=1}^{N} \frac{1}{N} \log_2 p(x_i)}$$

- ✓ SYNAT: Interdyscyplinarny system interaktywnej informacji naukowej i naukowo technicznej: transliteracja rzeczywistych nagrań wiadomości radiowych

Dekoder	ACC	CORR	PP
Autorski System LSTM	65	67	246
Julius + SRILM mix 40k	61	65	
Julius + IRSTLM 30K	44	52	289
Julius + IRSTLM 60K	48	58	276

Czym dysponujemy 1

- ✓ Moduł wyszukiwania mowy w sygnale (VAD)
  - ✓ MFCC, sieć LSTM, wyjście w formacie TextGrid (Praat)
  - ✓ Wysoka skuteczność, także w szumie, obecności muzyki
  - ✓ 0,1 czasu rzeczywistego
- ✓ Normalizator tekstu
  - ✓ Rozwijanie liczebników i skrótów
  - ✓ Uzgadnianie form gramatycznych liczebników i skrótów
  - ✓ Kombinacja LM dla słów, form podstawowych i POS, synchroniczny dekoder Viterbiego (ISMIS 2012)
- ✓ Koneksjonistyczny model języka
  - ✓ Sieć neuronowa modeluje zależności pomiędzy słowami
  - ✓ W eksperymentach perplexity mniejsze o ok. 20% niż najlepszy model n-gramowy (ISMIS 2011)

Czym dysponujemy 2

- ✓ Własny dekoder LSTM
  - ✓ autorska implementacja
  - ✓ sieci LSTM w modelowaniu akustycznym, koneksjonistyczny model języka
  - ✓ Zoptymalizowany, bardzo szybki - < czas rzeczywisty na laptopie
- ✓ Konwersja tekst ortograficzny – zapis fonetyczny
  - ✓ System regułowy
  - ✓ Warianty wymowy
- ✓ Spikes – rzadka reprezentacja sygnału
  - ✓ Kodowanie mowy w sposób zbliżony do działania słuchu
- ✓ Zasoby językowe
  - ✓ Nagrania i stenogramy senackie – część ręcznie poprawiona, ok. 300 h nagrań mowy, 5 mln słów (domenowy korpus transliteracji)
  - ✓ Teksty prawnicze z Wolters-Kluwer (ok. 1 mld słów)
  - ✓ Korpus nagrań radiowych 77 h nagrań mowy, transliteracja nagrań (727 tys. słów)
  - ✓ Zbieranie korpusów równoległych (Euronews)

Tłumaczenie mowy PL-EN

- ✓ EU-Bridge (www.eu-bridge.eu), FP 7, tłumaczenie mowy: napisy, parlament, aplikacje mobilne, wykłady
  - ✓ IWSLT 2012, pierwszy benchmark PL->EN, wykłady TED, *crowd sourcing* (tłumaczone przez wolontariuszy), truecase
  - ✓ dekoder Moses, training Giza++
  - ✓ BLEU jako miara jakości (porównanie tekstu MT z ręcznym tłumaczeniem)

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

  - ✓ Zbiór treningowy: ok. 130 tys. zdań, development: 767, test: 1564 (2 dodatkowe zbiory do oceny przez organizatorów)
  - ✓ Narzędzia: Politechnika Wrocławska, Stanford, własny tagger

BLEU	Słowo-słowo	Forma podstawowa PL	Faktor1: word:stem:tag	Faktor2: word:stem:tag
test	15,37	14,41	11,06	8,02
devel	20,36	18,74	13,22	9,47

U-STAR PL<->EN

- ✓ U-STAR consortium (NICT): 23 języki, rozproszona struktura serwerowa, standardy przyjęte przez ITU
- ✓ aplikacja VoiceTra4U na IPhone, IPad
- ✓ Równoczesna rozmowa do 5 osób równocześnie
- ✓ domena: rozmówki turystyczne
- ✓ PJWSTK: PL<->EN, na podstawie BTEC (110 tys. zdań, własne tłumaczenie, 2009), BLEU=71***
- ✓ ASR Julius, CentOS, Tomcat

Inne

- ✓ CLARIN – www.clarin.eu
- ✓ Synteza mowy polskiej – własny system syntezy korpusowej (Festival)
- ✓ Systemy dialogowe (Primespeech)
- ✓ Patologia mowy, jakość fonacji
- ✓ Zasoby mowy – ELRA





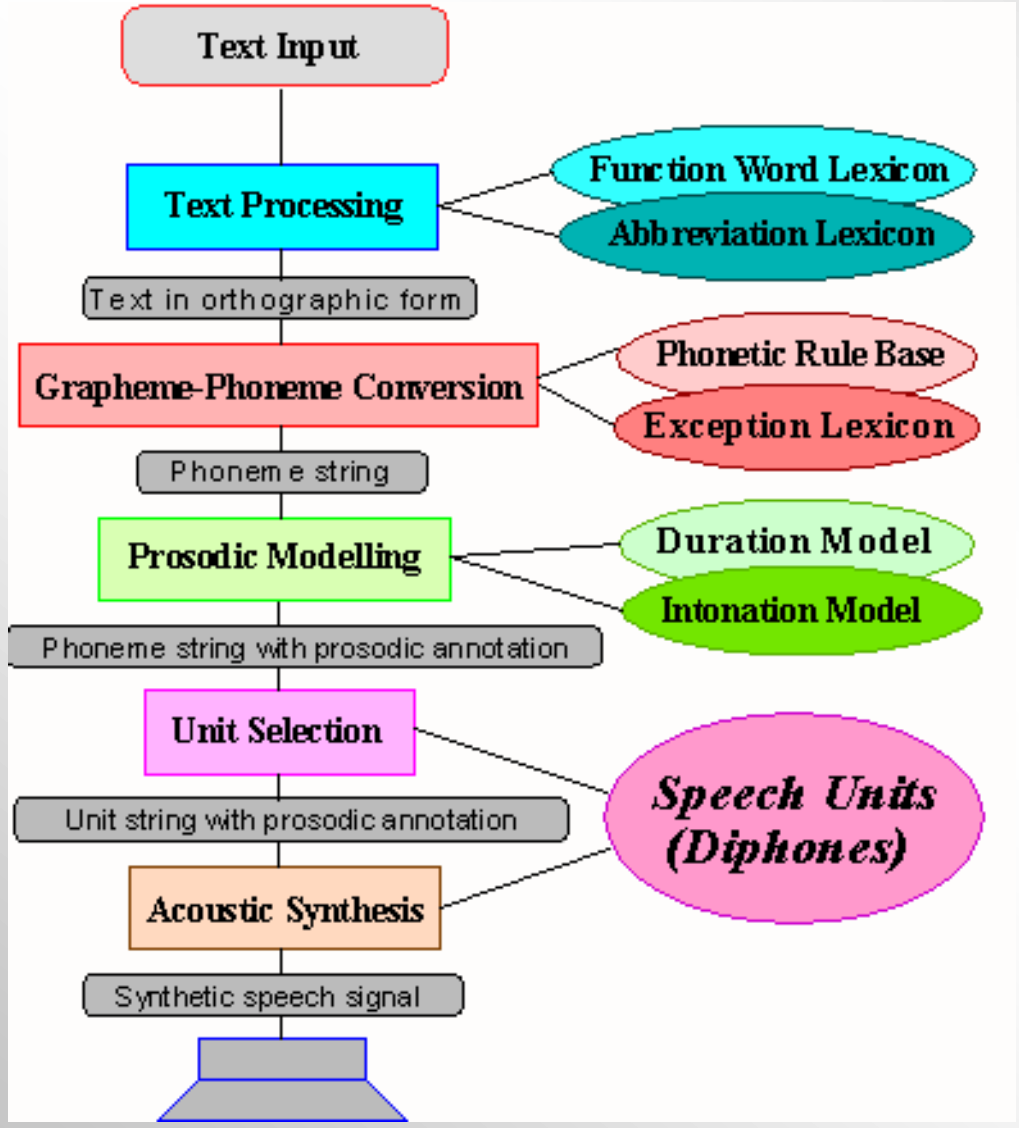

Multimedia Department

Dziękuje za uwagę!

☺

kmarasek@pjwstk.edu.pl

Synteza mowy

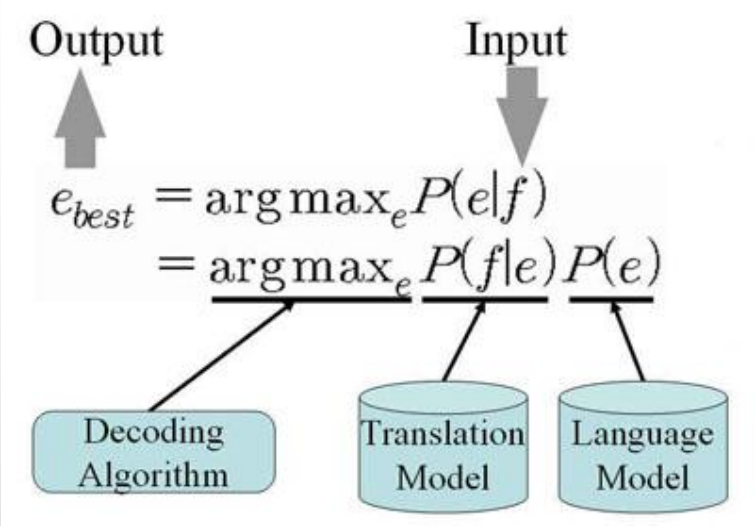


- ✓ regułowa
  - ✓ formantowa
  - ✓ artykulacyjna
- ✓ konkatenacyjna
  - ✓ difonowa
  - ✓ korpusowa

Multimedia Department

Automatyczne tłumaczenie statystyczne

- ✓ **Input :** SMT system otrzymuje zdanie do przetłumaczenia
- ✓ **Output :** SMT system generuje zdanie które jest tłumaczeniem zdania wejściowego
- ✓ **Language Model (model języka)** jest modelem określającym prawdopodobieństwo dowolnej sekwencji słów w danym języku
- ✓ **Translation Model (model tłumaczenia)** określa prawdopodobieństwa par tłumaczeń
- ✓ **Decoding Algorithm (dekodowanie**) to algorytm przeszukiwania grafu wyznaczający optymalne przejście przez graf słów

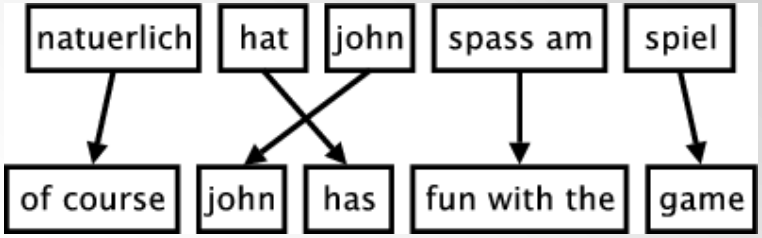


Jak to działa?

- ✓ coraz popularniejsze, coraz bardziej akceptowalne wyniki
  - ✓ Coraz rzadziej systemy regułowe, interlingua, modelowanie
  - ✓ Phrase based translation zamiast analizy składni, pasowania zdań
  - ✓ Każda fraza jest tłumaczona na frazę i wynik może mieć zmienioną kolejność
  - ✓ Model matematyczny

$\text{argmax}_{\mathbf{e}}\ p(\mathbf{e}|\mathbf{f}) = \text{argmax}_{\mathbf{e}}\ p(\mathbf{f}|\mathbf{e})\ p(\mathbf{e})$

model języka **e** i model tłumaczenia $p(\mathbf{f}|\mathbf{e})$



  - ✓ Podczas dekodowania sekwencja słów F jest dzielona na sekwencje *I* fraz f_1^I o jednakowym prawdopodobieństwie, każda z fraz f_i z f_1^I jest tłumaczona na frazę e_i i mogą one zostać poprzestawiane.
  - ✓ Tłumaczenie jest modelowane jako rozkład $\varphi(f_i|e_i)$, przestawianie fraz jako $d(start_i,end_{i-1}) = \alpha^{|start_i - end_{i-1} - 1|}$ z odpowiednią wartością α, wprowadza się też czynnik ω>1 aby chętniej generować krótsze zdania

Zatem najlepsze tłumaczenie to

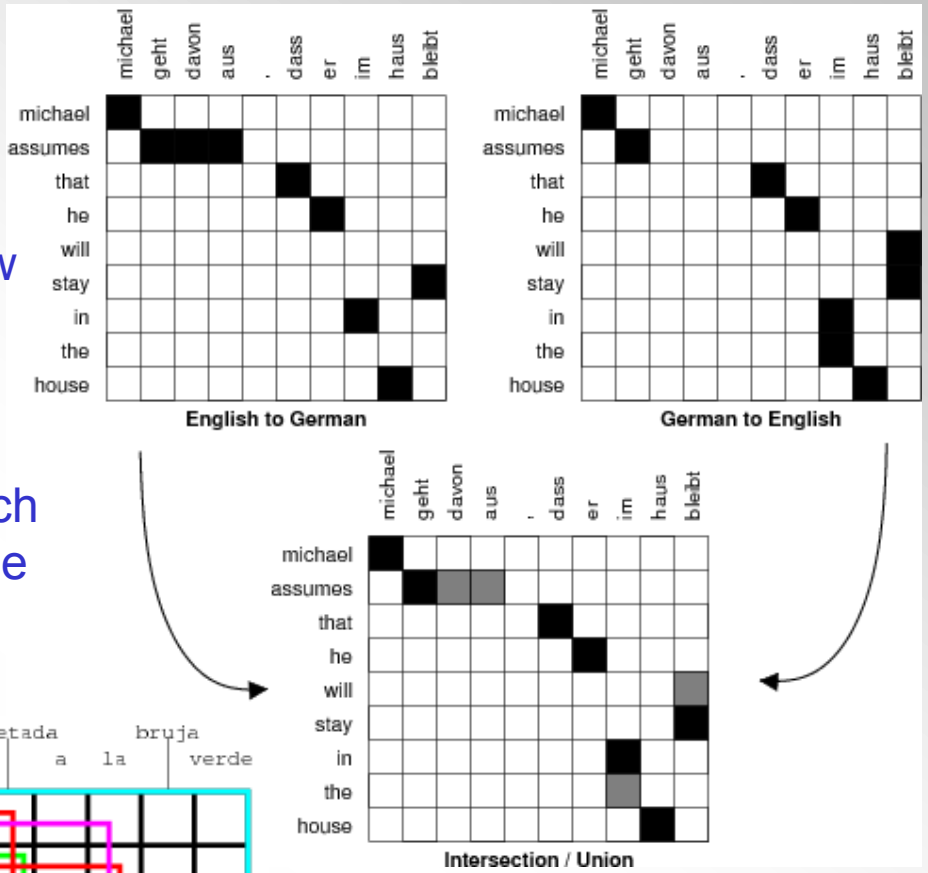
$\mathbf{e}_{best} = \text{argmax}_{\mathbf{e}}\ p(\mathbf{e}|\mathbf{f}) = \text{argmax}_{\mathbf{e}}\ p(\mathbf{f}|\mathbf{e})\ p\_LM(\mathbf{e})\ \omega^{length(\mathbf{e})}$ gdzie $p(\mathbf{f}|\mathbf{e})$ jest przedstawianie jako:

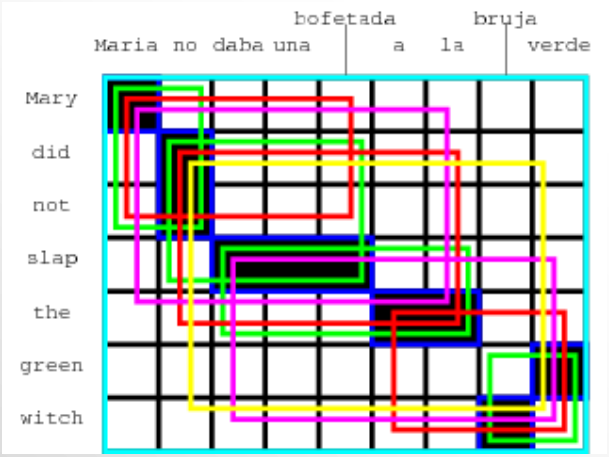
$p(f_1^I|e_1^I) = \Phi_{i=1}^I\ \varphi(f_i|e_i)\ d(start_i,end_{i-1})$

www.statmt.org/moses, 06

Tłumaczenie fraz, czyli phrase alignment

- ✓ Dopasowanie na podstawie bilingualnego korpusu – modelowanie
  - ✓ Dwukierunkowe dopasowanie słów pozwala wyznaczyć te najbardziej odpowiednie słowa
  - ✓ Na tej podstawie próbuje się tłumaczyć frazy, np. szukamy takich fraz które są dopasowane do siebie z wykorzystaniem tylko słów z ich wnętrza (Och, 2003)

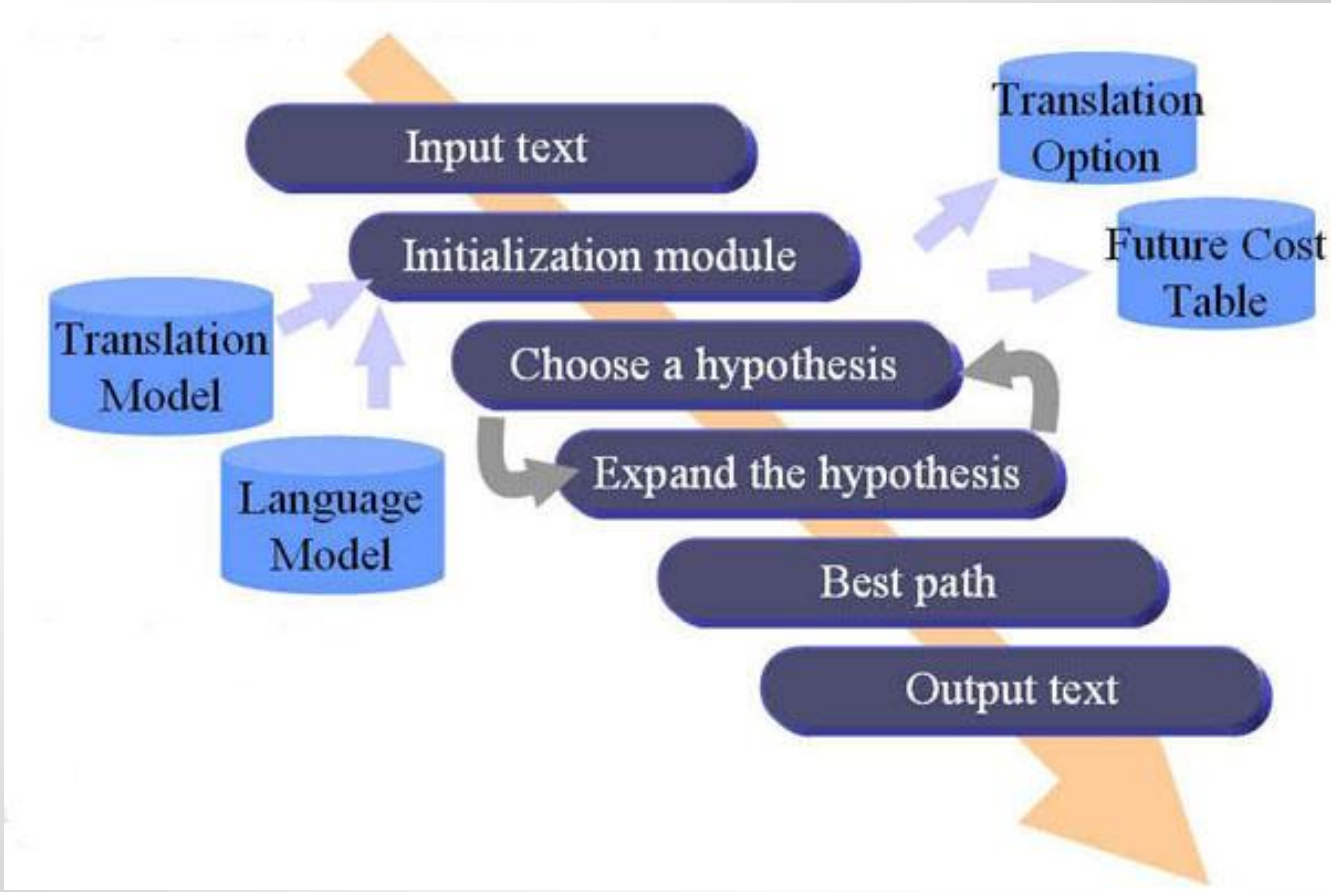




www.statmt.org/moses, 06

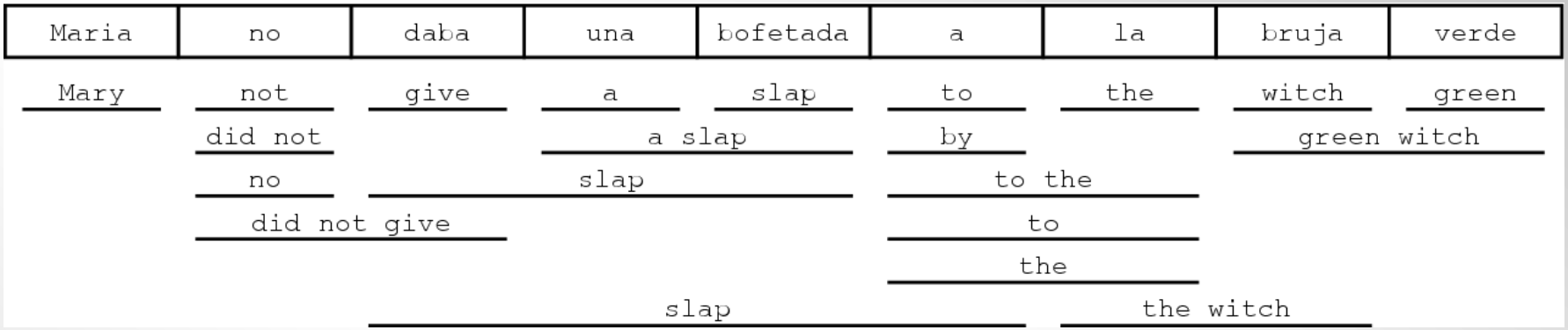
Multimedia Department

Proces dekodowania

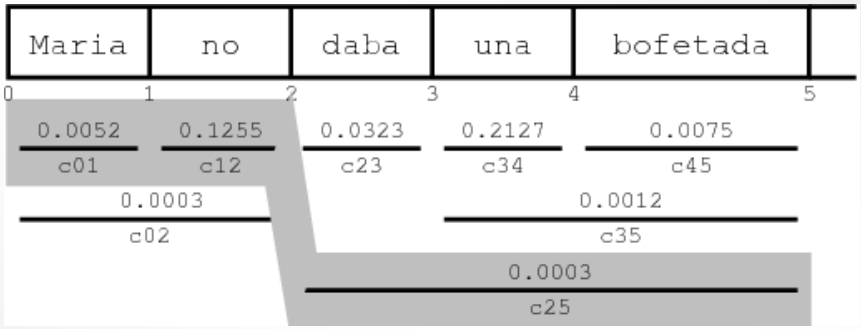


Dekodowanie czyli tłumaczenie

- ✓ System wyszukuje możliwe tłumaczenia fraz



- ✓ Uwzględnia koszt związany z tłumaczeniem, przestawianiem fraz i modelem języka – minimalizuje koszt, czyli maksymalizuje prawdopodobieństwo
- ✓ Metoda wyszukiwania jest zbliżona do używanej w rozpoznawaniu mowy

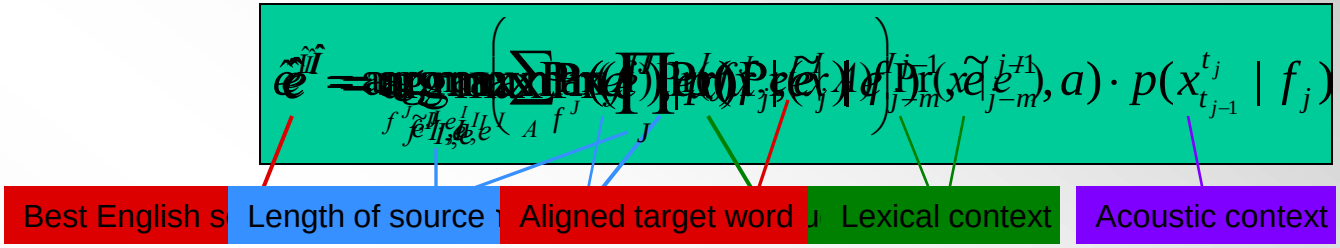


- ✓ Można generować listę najlepszych tłumaczeń

www.statmt.org/moses, 06

Sprzężenie rozpoznawania i tłumaczenia

- ✓ Wykorzystanie krat słów jako wejścia do tłumaczania z użyciem FST
- ✓ Noisy Channel Model

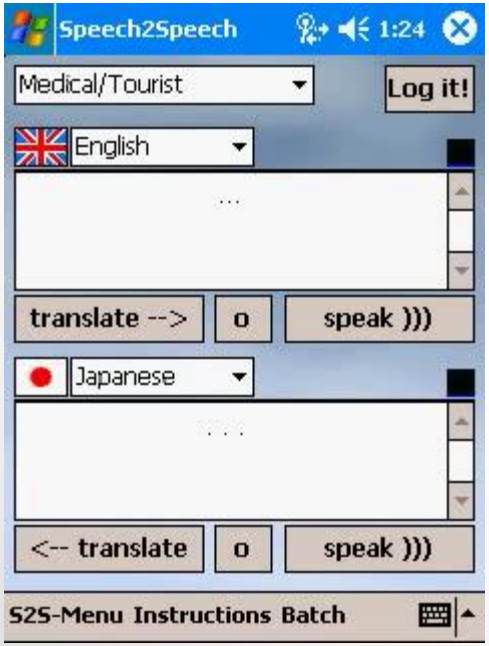


  - ✓ Using an alignment model, A
  - ✓ Instead of modeling the alignment, search for the best alignment

- ✓ Wykorzystanie interlingua i modeli sematycznych
  - ✓ Wykorzystanie prozodii do wspomagania parsingu zdań

Matusov, 05

Przykłady zastosowań



PDA

szkolenie

Multimedia Department

Przygotowanie korpusu BTEC dla języka polskiego

- ✓ Współpraca z ATR Spoken Language Translation Research Laboratories, Kyoto, dr Eiichiro Sumita
- ✓ Ok. 120 000 zdań z rozmów japońskich turystów mówiących po angielsku („angielskie rozmówki”)
- ✓ Cel:
  - ✓ Przygotowanie systemu automatycznego tłumaczenia z mowy na mowę pomiędzy polskim i angielskim dla ograniczonej domeny (rozmówki turystyczne)
  - ✓ Przygotowanie równoległego korpusu polsko-angielskiego (dopasowanie na poziomie zdań i fraz)
- ✓ Nasza praca:
  - ✓ Przetłumaczenie zdań z angielskiego na polski
  - ✓ Sformatowanie plików zgodnie z formatem BTEC
  - ✓ Anotacja morfo-syntaktyczna polskiego korpusu
- ✓ Aktualny stan zaawansowana:
  - ✓ Przetłumaczono ok. 80% tekstów

ATR Advanced Telecommunications Research Institute International

SLC 音声言語コミュニケーション研究所 Spoken Language Communication Research Laboratories